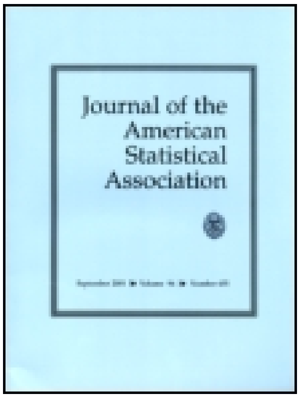

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC

Eun Sug Park[a], Peter Guttorp[a] & Ronald C Henry[a]

[a] Eun Sug Park is Assistant Research Scientist, Texas Transportation Institute, The Texas A&M University System, College Station, TX 77843. Peter Guttorp is Professor of Statistics and Director of the National Research Center for Statistics and the Environment, University of Washington, Seattle, WA 98195. Ronald C. Henry is Associate Professor of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA 90089. This research was conducted when the first author was Research Associate at National Research Center for Statistics and the Environment of University of Washington. Although the research described in this article has been funded by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. The authors gratefully acknowledge helpful comments from the reviewers.

Published online: 31 Dec 2011.

To cite this article: Eun Sug Park, Peter Guttorp & Ronald C Henry (2001) Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC, Journal of the American Statistical Association, 96:456, 1171-1183, DOI: 10.1198/016214501753381823

To link to this article: http://dx.doi.org/10.1198/016214501753381823

Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC

Eun Sug Park, Peter Guttorp, and Ronald C. Henry

Multivariate receptor modeling aims to estimate pollution source profiles and the amounts of pollution based on a series of ambient concentrations of multiple chemical species over time. Air pollution data often show temporal dependence due to meteorology and/or background sources. Previous approaches to receptor modeling do not incorporate this dependence. We model dependence in the data using a time series approach so that we can incorporate extra sources of variability in parameter estimation and uncertainty estimation. We estimate parameters using the Markov chain Monte Carlo method, which makes simultaneous estimation of parameters and uncertainties possible. The methods are applied to simulated data and 1990 Atlanta air pollution data. The results show promise towards the goal of accounting for the dependence in the data.

KEY WORDS: Air pollution; Chemical species; Compositions; Dynamic models; Gibbs sampler; Kalman filter; Metropolis–Hastings algorithm; Source profile.

1. INTRODUCTION

An important problem in environmental statistics is to determine the main sources of air pollution from data obtained at a given station, or receptor. To do so, data need to contain observations on the amounts (concentrations) of different chemical compounds, or species, in the atmosphere that are received (measured) at the station. Samples of airborne pollution are subjected to extensive chemical analysis. Contributing sources leave chemical fingerprints in the sample. The amount of pollution coming from each source can be estimated if the chemical fingerprints of the sources are known. This subfield of environmental statistics is called receptor modeling.

In this article, the pollutants of concern are volatile organic compounds (VOC) observed in downtown Atlanta, GA during July and August of 1990. VOCs are important because some, such as benzene and toluene, are toxic and many react with nitrogen oxides in the air to form ozone, which can reach dangerous levels in many areas. The chemical species of interest are predominately hydrocarbons containing from 2 to 10 carbon atoms. Methane with one carbon is excluded because of its large background concentration from ubiquitous natural sources and the fact that methane is of no interest to air pollution regulators. For this reason, the total amount of VOCs in the air is commonly known as total nonmethane organic compounds (TNMOC). The VOCs in this study were determined by an innovative automated gas chromatograph (GC) that sampled the air for 50 minutes each hour and analyzed the sample in the remainder of the hour (Purdue 1991). The detector of the GC responds to the number of carbon atoms in the sample, therefore the units are in parts per billion by volume of carbon (ppbC). The mass of the VOCs in this study is dominated by carbon, so these units are proportional to mass units such as nanograms per cubic meter used for airborne particulate matter. Of the more than 1,200 hourly observations of more than 40 VOCs, a dataset of 538 observations on 38 species (37 VOC and TNMOC) measured at one location was selected (Henry, Lewis, and Collins 1994). The background concentrations of these VOCs are small compared to the very large sources in the urban center of Atlanta and can be ignored. These major sources of VOCs are three in number and all are related to gasoline and diesel fueled vehicles. Obviously, there is the tailpipe exhaust of the vehicles when being driven. However, vehicles also emit VOCs by evaporation when sitting still and when running. These evaporative emissions can be classified as two sources, one source with the composition of whole gasoline, and another source with the composition of gasoline headspace vapor. Headspace vapor is the vapor above gasoline in a container, such as a fuel tank. This vapor is enriched in the more volatile compounds in gasoline, e.g. n-butane and isopentane. Diesel fuel is much less volatile than gasoline and evaporative emissions of diesel-fueled vehicles are negligible.

The assumptions of receptor modeling require conservative species. The species in the analysis must be conserved (in a relative sense, that is, with respect to the other species) during transport between the source and the receptor because, as shown later, the fundamental equations assume mass balance (see, e.g., Hopke 1985, 1991, 1997). Virtually all the VOCs in the dataset are reactive in the atmosphere and thus do not strictly obey mass conservation. The receptor modeling in this article is restricted to species which have atmospheric lifetimes (based on reaction rates with the hydroxyl radical) that are long compared to the travel time between source and receptor, which in this case is less than 2 hours. Thus, these species can be considered to be unreactive because their rate of reaction is too slow to significantly alter their concentrations between the source and the sampling site. Inclusion of highly reactive species will distort the source fingerprints to the extent that the respective source types in the ambient data cannot be identified. Also, in many environmental applications, some species have a few common major sources and some have many more

Eun Sug Park is Assistant Research Scientist, Texas Transportation Institute, The Texas A&M University System, College Station, TX 77843 (E-mail: *e-park@ttimail.tamu.edu*). Peter Guttorp is Professor of Statistics and Director of the National Research Center for Statistics and the Environment, University of Washington, Seattle, WA 98195. Ronald C. Henry is Associate Professor of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA 90089. This research was conducted when the first author was Research Associate at National Research Center for Statistics and the Environment of University of Washington. Although the research described in this article has been funded by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. The authors gratefully acknowledge helpful comments from the reviewers.

Journal of the American Statistical Association
December 2001, Vol. 96, No. 456, Applications and Case Studies

minuscule sources. We are concerned with the major pollution sources, not all sources since there could be hundreds of sources in nature, and it would be impossible or meaningless if we would try to identify all of those sources. Owing to these limitations (chemical reactivity and the number of pollution sources), the first important step in multivariate receptor modeling is to select an appropriate subset of species for an analysis though we do not pursue this issue in this article. By ‘appropriate’, we mean, ‘relatively unreactive species contributed by major pollution sources’. Because our goal is to model the three vehicle-related source emissions that are dominant in downtown Atlanta, the species related to other sources are also excluded. For instance, Toluene is excluded from the analysis because it is contributed by an additional solvent source as well as by the three vehicle-related sources, and inclusion of it would change the number of sources to be modeled. Nine vehicle-related species (see Table 1) are thus selected out of 38 species based on an environmental expert’s judgment, resulting in a final dataset of 538 observations on nine variables (VOCs). These particular species have lifetimes in the atmosphere that are long (greater than 12 hours) compared to the time scale of the observations and transport times from source to receptor, which for our data are both about 1 hour.

Traditionally, there have been two different approaches in receptor modeling, the chemical mass balance (CMB) receptor model and the multivariate receptor model, depending on whether the chemical fingerprints of the sources are assumed known or not (see Henry, Lewis, and Hopke 1984, Henry 1991). In receptor modeling terms, the source fingerprint is often called the source composition profile. It consists of the relative amount of each chemical species in the emissions from the source, and so is unit-free. In many cases, source composition profiles are unavailable either because we do not know the contributing sources, or because direct measurement of source emissions is very difficult and expensive (e.g., for mobile sources). Even in the case where direct source measurement is available, it still may not be representative of the source emissions in the airshed under consideration (e.g., the distribution of vehicles for tunnel measurements of autoemissions may not be the same as that which is contributing most heavily at the receptor). Pollutant transport, reactions, measurement errors, variations in source compositions, and the contribution of minor sources can also make the deviation larger between the measured source profiles and the true source compositions for the ambient data.

In Atlanta, there were also three vehicle-related profiles obtained by direct source measurements during the summer of 1990: highway tunnel measurements, whole gasoline, and gasoline headspace (Conner, Lonneman, and Seila 1995; Henry et al. 1994), which may allow an application of CMB model (assuming those measured source compositions are not biased). They are given in Table 1 for nine selected vehicle-related VOC species. It should, however, be noted that those direct source measurements were obtained, under rather restricted conditions, independently of the data (e.g., roadway compositions were obtained as highway tunnel measurements during morning rush hour). In this article, we estimate the compositions for the three vehicle-related sources in downtown Atlanta; roadway emissions, whole gasoline, and gasoline headspace vapor from ambient data using a multivariate receptor model. This will make an objective comparison between the estimated source compositions based on the ambient data and the measured source compositions possible.

The chemical mass balance equation is the physical basis for most receptor models (both CMB models and multivariate receptor models). Mathematically, it can be written (after inclusion of error terms) as follows:

$$y_t = \sum_{k=1}^{q} \alpha_{tk} P_k + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{1}$$

Here, $y_t = (y_{t1}, y_{t2}, \ldots, y_{tp})$ is the measured concentrations of p chemical species at time t, q is the number of sources, $P_k = (p_{k1}, p_{k2}, \ldots, p_{kp})$ is the source composition profile (source fingerprint) for source k, α_{tk} is the contribution from source k in time t, and is $\varepsilon_t = (\varepsilon_{t1}, \varepsilon_{t2}, \ldots, \varepsilon_{tp})$ is the measurement error associated with y_t. As already mentioned, the source composition profiles P_k $(k = 1, \ldots, q)$ are assumed known in CMB models, whereas they are the unknown key parameters in multivariate receptor models. In matrix terms, the model (1) can be written as

$$Y = AP + E, \tag{2}$$

where A is $n \times q$ source contribution matrix, P is $q \times p$ source composition matrix, and E is $n \times p$ error matrix.

The assumption of independence among the observations y_t has been made either implicitly or explicitly in all previous approaches to receptor modeling, see, for instance, Hopke (1991), Henry (1991), Yang (1994), Gleser (1997), Park (1997), and Park, Spiegelman, and Henry (2001). Air pollution data, however, are usually obtained as a series of measurements on concentrations of aerosols over time, and meteorology often induces some degree of dependence in the data. Observations closer in time tend to be more correlated than observations farther apart in time. In some cases the assumption of independence may not be grossly wrong because environmental data usually contain many missing values or erroneous observations, and after initial screening of the data, time separation between any pair of measurements may become

Table 1. Measured Source Composition Profiles in Atlanta

Source	*acetylene*	*propene*	*nButane*	*2MPentan*	*3MPentan*	*benzene*	*CyHx+2MHx*	*2,3-DMP*	*2,2,4-TMP*
Roadway	.181	.094	.197	.116	.069	.132	.049	.043	.120
Gasoline	0	.002	.197	.221	.138	.108	.116	.067	.152
Headspace	0	.007	.685	.144	.075	.034	.021	.014	.021

NOTE: Each source profile is normalized to sum to one.

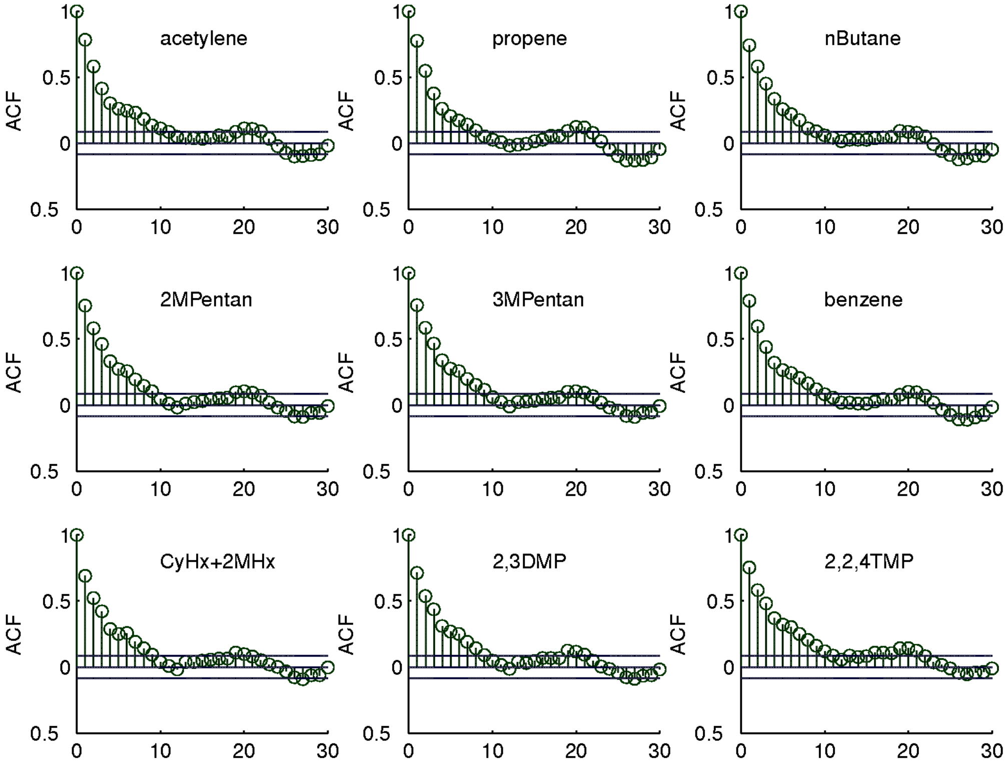

Figure 1. Autocorrelation Function (ACF) Plots of Y for the Atlanta Data.

large enough so that serial correlation can be ignored in the screened data. This, of course, is not always the case. Figure 1 contains the autocorrelation function (ACF) plots for nine selected vehicle-related species for the Atlanta data.

Assuming that the measured compositions in Table 1 are the true source compositions, that is, P is known in model (2), A can be estimated easily, for instance, as an ordinary least squares (OLS) solution, $\widehat{A}_{\mathrm{OLS}} = YP'(PP')^{-1}$, if we ignore the dependence structure in the data and vice versa, that is, $\widehat{P}_{\mathrm{OLS}} = (A'A)^{-1}A'Y$ if A is known or estimated first. This was done in almost all previous work without checking the independence assumption. Figure 2 shows the ACF plots of the residuals calculated as $Y - \widehat{A}_{\mathrm{OLS}}P$ for each of nine species, and Figure 3 shows ACF plots of OLS estimates of source contributions, $\widehat{A}_{\mathrm{OLS}}$.

All three plots, Figures 1–3, reveal significant serial correlation in the data. It is well known in time series literature that in the presence of the correlated residuals, the standard error (not adjusting for the correlation in the residuals) of OLS estimate of the trend (which may be regarded as P in our model) in the regression is often grossly wrong. Although the correct standard error of OLS estimate may be obtained by adjusting for the correlation, it is still not the best estimate since the generalized least squares estimate, taking the correlation into account in the estimation procedure, has smaller standard error.

The goal of this article is to extend multivariate receptor models to account for temporal dependence in the data so that we can incorporate that source of variability into the estimation of parameters and uncertainties. In Section 2, we introduce models accounting for time dependence in the observations. Estimation of parameters is discussed in Section 3. Sections 4 and 5 contain examples from simulated data and the Atlanta air pollution data, respectively. Finally, concluding remarks are made in Section 6.

2. MODEL

The model (1) may be viewed as a factor analysis model in the sense that Y is the only observable quantity whereas q (number of factors), P (factor loading matrix), and A (factor score matrix) are all unknown quantities that need to be estimated (or predicted). Although estimation of q is not a trivial problem, it is not the purpose of this article, and so it is assumed as known throughout this article.

It is well-known that, without imposing additional constraints on the parameters, the factor analysis model is not identifiable even with known number of sources, q. There have been several attempts to avoid nonidentifiability of the factor analysis model in multivariate receptor modeling by imposing more restrictive constraints on either the P, or the A matrix (see Henry and Kim 1990; Henry et al. 1994; Yang 1994; Henry 1997; Park 1997; Park, Spiegelman, and Henry 2001). These additional constraints on the parameters are called “identifiability conditions.” As a matter of fact, there could be many different sets of identifiability conditions, each making sense in its own context (see Park et al. 2001 for some identifiability conditions that are meaningful in receptor models). Here, we borrow conditions from the confirmatory factor analysis model (see e.g., Anderson 1984), which is often realistic

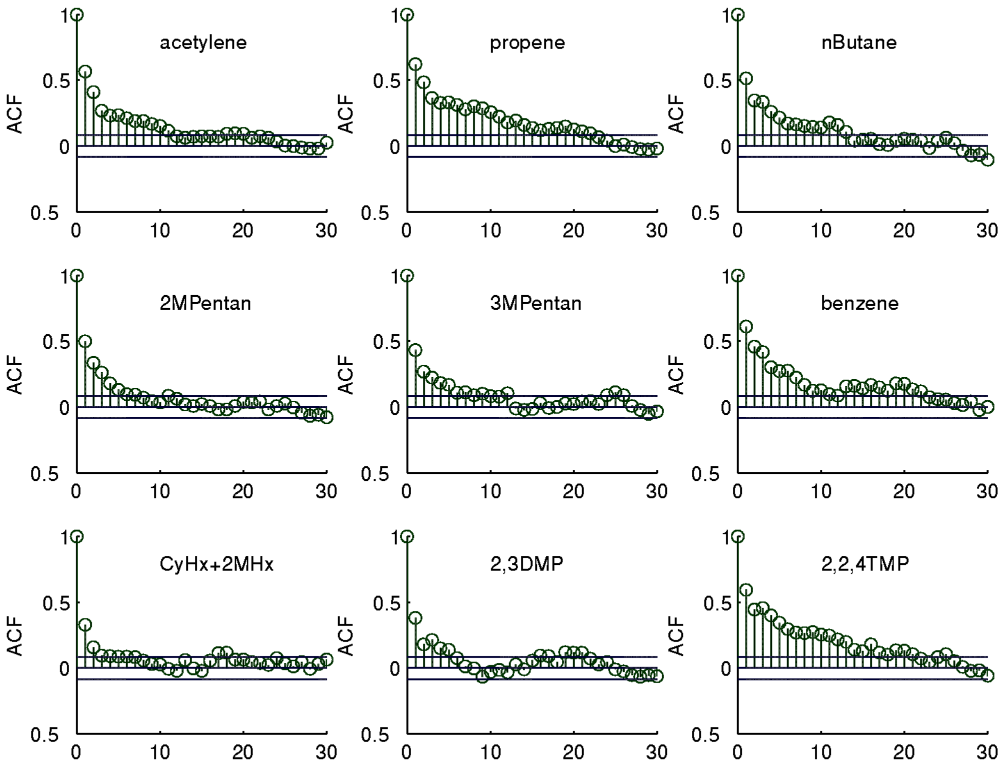

Figure 2. Autocorrelation Function (ACF) Plots of the Residuals ($Y - \widehat{A}_{OLS}P$, where P is the measured source compositions in Table 1) for the Atlanta Data.

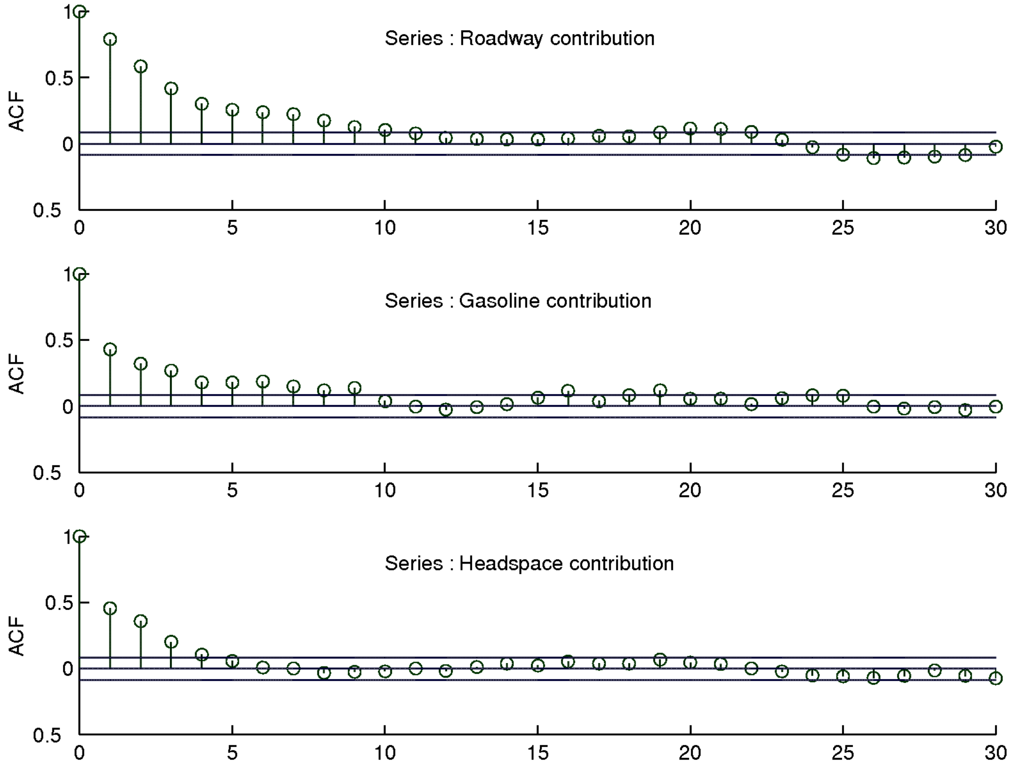

Figure 3. Autocorrelation Function (ACF) Plots of Source Contributions ($\widehat{A}_{OLS}$) for the Atlanta Data.

in air pollution studies. The assumptions are:

C1. There are at least $q-1$ zero elements in each row of P.

C2. The rank of $P^{(k)}$ is $q-1$, where $P^{(k)}$ is the matrix composed of the columns containing the assigned 0s in the kth row with those assigned 0s deleted.

In terms of air pollution data, C1 implies that some pollutants (corresponding to zeros) are not contributed by a particular source, and C2 implies that no two sources share the same set of zeros. This set of assumptions is weaker than the tracer element assumption requiring that each source contributes only once, which was traditionally made in receptor modeling as a way of resolving the nonidentifiability problem (see Park et al. 2001). Under the above conditions, the source profiles, P, are identified up to normalization, which is enough for the purpose of a receptor model: as long as the relative amount of each species in a source is determined, a source can be identified.

Thus, our analysis in this article is conditional on the model that assumes a known number of sources and predetermined identifiability conditions, and focuses on developing a factor analysis model for temporally correlated observations. Alternatively, Park, Oh, and Guttorp (2000) discusses the problem of model uncertainty in multivariate receptor modeling by treating q and identifiability conditions as unknown under the assumption that the observations are independent. A Bayesian approach is used in Park et al. (2000), considering the posterior probabilities for a range of plausible models obtained by varying q and identifiability conditions.

Assume that the y_t in (1) are dependent. We first need to decide how to model this dependence. It seems reasonable to assume that the source contribution at time t depends on the past source contributions (as Figure 3 indicates). Also, it is often the case that ε_t contains not only pure measurement error but also all the remaining sources of variability that are not explained by the systematic part of our model, such as background sources (unmodeled minor sources), variations in source compositions, and meteorology. Then it is likely that the ε_t are also correlated in time due to the effect of those (see Figure 2). We may decompose ε_t into two terms $\varepsilon_t = \eta_t + \delta_t$, where η_t represents variability correlated in time owing to meteorology or background sources, and δ_t represents residual, unpredictable variability owing to pure measurement error, independent over time.

We consider the model

$$y_t = \alpha_t P + \eta_t + \delta_t,$$

where $\alpha_t = (\alpha_{t1}, \alpha_{t2}, \ldots, \alpha_{tq})$ is a stationary vector AR(1) process centered at $\xi = (\xi_1, \xi_2, \ldots, \xi_q)$, $\eta_t = (\eta_{t1}, \eta_{t2}, \ldots, \eta_{tp})$ is a stationary vector AR(1) process centered at $\mathbf{0}$, and $\delta_t = (\delta_{t1}, \delta_{t2}, \ldots, \delta_{tp}) \sim N_p(\mathbf{0}, \Sigma)$ where $\Sigma = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. We use '$N_k(\cdot,\cdot)$' to denote k-dimensional multivariate normal distribution throughout this article. We also assume that the elements of P are all nonnegative because the source composition cannot contain any negative values. This model may be written in dynamic linear model (DLM) form (West and Harrison, 1997) as

Observation equation: $y_t = \alpha_t P + \eta_t + \delta_t, \quad \delta_t \sim N_p(\mathbf{0}, \Sigma),$

State equation:

$$\begin{aligned} \alpha_t &= \xi + (\alpha_{t-1} - \xi)\Phi + u_t, \quad u_t \sim N_q(\mathbf{0}, U), \\ \eta_t &= \eta_{t-1}\Theta + v_t, \quad v_t \sim N_p(\mathbf{0}, V), \end{aligned} \tag{3}$$

where $u_t = (u_{t1}, u_{t2}, \ldots, u_{tq})$, $\Phi = \operatorname{diag}(\phi_1, \phi_2, \ldots, \phi_q)$, ϕ_k is an AR coefficient for the kth source contribution, $v_t = (v_{t1}, v_{t2}, \ldots, v_{tp})$, $\Theta = \operatorname{diag}(\theta_1, \theta_2, \ldots, \theta_p)$, and θ_j is an AR coefficient for jth element of η_t. Note that marginal distribution for each α_t is

$$\alpha_t \sim N_q(\xi, W), \quad W = \Phi W \Phi + U \tag{4}$$

and for each η_t is

$$\eta_t \sim N_p(\mathbf{0}, M), \quad M = \Theta M \Theta + V. \tag{5}$$

Equivalently, model (3) can be reparameterized in terms of the centered source contributions. Let $\gamma_t = \alpha_t - \xi$ and $\mu = \xi P$. Then we have

Observation equation:

$$y_t = \mu + \gamma_t P + \eta_t + \delta_t, \quad \delta_t \sim N_p(\mathbf{0}, \Sigma),$$

State equation:

$$\begin{aligned} \gamma_t &= \gamma_{t-1}\Phi + u_t, \quad u_t \sim N_q(\mathbf{0}, U), \\ \eta_t &= \eta_{t-1}\Theta + v_t, \quad v_t \sim N_p(\mathbf{0}, V). \end{aligned} \tag{6}$$

The marginal distribution for each γ_t is

$$\gamma_t \sim N_q(\mathbf{0}, W), \quad W = \Phi W \Phi + U \tag{7}$$

and for each η_t remains the same as (5). Recall that under identifiability conditions C1 and C2, P is identified up to a normalizing constant, as is the mean contribution ξ. In other words, ξ is not identified in model (3) unless we have extra information such as the total mass of pollutant particles (a normalizing constant for P). For this reason, we proceed with the parameterization (6) here. However, it should be noted that the parameterization (3) might be preferred if we have additional information to remove unidentifiability of ξ.

3. ESTIMATION

As the model gets complicated by the inclusion of more parameters, Markov chain Monte Carlo (MCMC) simulation (Tierney 1994; Chib and Greenberg 1995; Besag, Green, Higdon, and Mengersen 1995; Gilks, Richardson, and Spiegelhalter 1996) seems to be an attractive approach for parameter estimation. Also note that the parameters of the models (3) or (6) are all unknown, and the problem of parameter estimation is essentially nonlinear, but the MCMC method makes the problem linear by use of conditional distributions. We employ an MCMC method as a computational tool under a Bayesian framework. As mentioned in Section 2, a multivariate receptor model can be viewed as a special type of a factor analysis model with the constraints that the elements of factor loading matrix P should all be nonnegative. These nonnegativity constraints and model identifiability conditions C1 and C2 can be absorbed into the prior distribution for P.

Under the normal error assumption on δ, the likelihood $f(Y|\cdots)$ is written as

$$f(Y|\cdots) = |2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}tr\Sigma^{-1}\sum_{t=1}^{n}(y_t-\mu-\gamma_t P-\eta_t)' \times (y_t-\mu-\gamma_t P-\eta_t)\right\}. \qquad (8)$$

We use '$|\cdots$' to denote conditioning on all other variables. For the sake of brevity, μ is assumed known as it is an incidental parameter here. For a prior distribution $p(\cdot)$, we assume that

$$\begin{aligned} p(P, \Sigma, \Phi, U, \gamma_1, \ldots, \gamma_n, \Theta, V, \eta_1, \ldots, \eta_n) &= p(P)p(\Sigma)p(\Phi)p(U)p(\gamma_1, \ldots, \gamma_n|\Phi, U) \\ &\quad \times p(\Theta)p(V)p(\eta_1, \ldots, \eta_n|\Theta, V). \end{aligned}$$

Note that (6) implies

$$\begin{aligned} p(\gamma_1, \ldots, \gamma_n|\Phi, U) &= (2\pi)^{-\frac{n}{2}}|W|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\gamma_1 W^{-1}\gamma_1'\right)|U|^{-\frac{n-1}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2}trU^{-1}\sum_{t=2}^{n}(\gamma_t-\gamma_{t-1}\Phi)'(\gamma_t-\gamma_{t-1}\Phi)\right\} \end{aligned}$$

and

$$\begin{aligned} p(\eta_1, \ldots, \eta_n|\Theta, V) &= (2\pi)^{-\frac{n}{2}}|M|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\eta_1 M^{-1}\eta_1'\right)|V|^{-\frac{n-1}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2}trV^{-1}\sum_{t=2}^{n}(\eta_t-\eta_{t-1}\Theta)'(\eta_t-\eta_{t-1}\Theta)\right\}. \end{aligned}$$

Based on a series of observations $y_1, \ldots, y_n$, we are interested in sampling from the full posterior $\pi(P, \Sigma, \Phi, U, \gamma_1, \ldots, \gamma_n, \Theta, V, \eta_1, \ldots, \eta_p \mid Y)$. We use "block-at-a-time" Metropolis–Hastings algorithm (Chib and Greenberg, 1995). We shall make use of seven move types in implementing MCMC:

(a) updating P,
(b) updating Σ,
(c) updating Φ,
(d) updating U,
(e) updating Θ,
(f) updating V,
(g) updating $\gamma_1, \ldots, \gamma_n$, and $\eta_1, \ldots, \eta_n$.

Let Γ be $n \times q$ matrix of which rows are γ_t, H be $n \times p$ matrix of which rows are η_t, $t = 1, \ldots, n$, and $\mathbf{1_n}$ be n-dimensional column vector of ones. Letting

$$\begin{aligned} \widetilde{P} &= (\Gamma'\Gamma)^{-1}\Gamma'(Y - \mathbf{1_n} \otimes \mu - \mathrm{H}), \\ S &= (Y - \mathbf{1_n} \otimes \mu - \mathrm{H} - \Gamma\widetilde{P})'(Y - \mathbf{1_n} \otimes \mu - \mathrm{H} - \Gamma\widetilde{P}), \end{aligned}$$

and using the orthogonality properties associated with $\widetilde{P}$ (see Press 1982), equation (8) can be written as

$$\begin{aligned} &|2\pi\Sigma|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}tr\Sigma^{-1}S\right) \\ &\quad \times \exp\left\{-\frac{1}{2}tr\Sigma^{-1}(P-\widetilde{P})'(\Gamma'\Gamma)(P-\widetilde{P})\right\} \\ &\quad \propto \exp\left\{-\frac{1}{2}(\mathrm{vec}\,P - \mathrm{vec}\,\widetilde{P})'(\Sigma^{-1}\otimes\Gamma'\Gamma)(\mathrm{vec}\,P - \mathrm{vec}\,\widetilde{P})\right\}. \end{aligned}$$

Let the prior distribution for P be

$$\begin{aligned} p(P) = p(\mathrm{vec}\,P) &\sim N_{pq}(c_0, C_0) \\ &\quad \times \mathbf{I}(P_{kj} \geq 0, \quad k = 1, \ldots, q, \quad j = 1, \ldots, p), \end{aligned}$$

where c_0 is a pq-dimensional vector and C_0 is a $pq \times pq$-dimensional diagonal matrix. Enforcing the constraints C1 and C2 is equivalent to using a degenerate point prior for some of the elements of P. We set $q \times (q-1)$ elements of c_0 and the corresponding elements of C_0 to be zero, which makes the prior distribution for P a truncated singular normal distribution (though still proper). Then the resulting full conditional posterior distribution $\pi(P|\cdots)$ is again a truncated singular normal distribution, which can be written as

$$\begin{aligned} \mathrm{vec}\,P|\cdots &\sim N_{pq}(c, C) \\ &\quad \times \mathbf{I}(P_{kj} \geq 0, \quad k = 1, \ldots, q, \quad j = 1, \ldots, p), \end{aligned}$$

where

$$\begin{aligned} c &= C\{(\Sigma^{-1}\otimes\Gamma')\mathrm{vec}\,(Y - \mathbf{1_n}\otimes\mu - \mathrm{H}) + C_0^- c_0\}, \\ C &= (\Sigma^{-1}\otimes\Gamma'\Gamma + C_0^-)^{-1}, \end{aligned}$$

where C_0^- is a generalized inverse of C_0. Since both Σ and C_0 are diagonal, samples of P can be obtained by sampling a subvector of vec P corresponding to each column of P separately as in Park et al. (2000), using a simple Metropolis–Hastings algorithm.

Under a usual inverse gamma, prior distribution for σ_j^2, $\sigma_j^{-2} \sim \mathrm{Gamma}(\alpha_{0j}, \beta_{0j})$, $j = 1, \ldots, p$, with the parameterization in which the mean and variance are α_{0j}/β_{0j} and α_{0j}/β_{0j}^2, respectivley, the full conditional for $\{\sigma_j^2\}$ is

$$\sigma_j^{-2}|\cdots \sim \mathrm{Gamma}\left(\alpha_{0j} + \frac{1}{2}n, \beta_{0j} + \frac{1}{2}d_j\right),$$

where d_j is the jth diagonal element of the matrix $(Y - \mathbf{1_n} \otimes \mu - \mathrm{H} - \Gamma P)'(Y - \mathbf{1_n} \otimes \mu - \mathrm{H} - \Gamma P)$. This can be easily sampled using a Gibbs sampler.

Moves (c)–(g) require Metropolis–Hastings steps. We use the same strategy as those given in Chib and Greenberg (1995) and West and Harrison (1997) to update Φ and U, respectively. Under uniform priors for ϕ_k, writing $\phi = (\phi_1, \ldots, \phi_q)$ for the diagonal of Φ, and $\Delta = \mathrm{diag}(\gamma_{t-1})$, the full conditional posterior density for Φ, $\pi(\phi|\cdots)$, is proportional to

$$g(\Phi)f_{\mathrm{nor}}(\phi|\tau, \mathrm{T})\,\mathbf{I}(0 < \phi < 1),$$

where f_{nor} is the q-variate normal density function, $\mathrm{T}^{-1} = \sum_{t=2}^{n} \Delta' U^{-1} \Delta, \tau = \mathrm{T} \sum_{t=2}^{n} \gamma_t U^{-1} \Delta', g(\Phi) = |W|^{-\frac{1}{2}} \exp(-\frac{1}{2}\gamma_1 W^{-1} \gamma_1')$, $W = \Phi W \Phi + U$, and $\mathbf{I}(0 < \phi < 1) = \prod_{k=1}^{q} \mathbf{I}(0 < \phi_k < 1)$. We use $N_q(\tau, \mathrm{T})$ as a proposal distribution for ϕ (independent proposal) and accept the proposal ϕ^* with probability

$$\min\left\{1, \frac{g(\Phi^*)\,\mathbf{I}(\mathbf{0} < \phi^* < \mathbf{1})}{g(\Phi)\,\mathbf{I}(\mathbf{0} < \phi < \mathbf{1})}\right\},$$

where $W^* = \Phi^* W^* \Phi^* + U$.

The full conditional posterior for U, $\pi(U|\cdots)$, is proportional to

$$p(U) g(U) |U|^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2} tr U^{-1} B\right),$$

where $B = \sum_{t=2}^{n} (\gamma_t - \gamma_{t-1}\Phi)'(\gamma_t - \gamma_{t-1}\Phi)$ and $g(U) = |W|^{-\frac{1}{2}} \exp(-\frac{1}{2}\gamma_1 W^{-1}\gamma_1')$. Note that B follows a Wishart distribution with parameters U and $n-1$, $B \sim W(U, n-1)$. Under an inverted Wishart prior $U \sim W^{-1}(\Psi_0, r_0)$, the conditional distribution of U given B is $U|B \sim W^{-1}(\Psi_0 + B, r_0 + n - 1)$. Here we use the parameterizations of the Wishart and inverted Wishart distributions as given in Anderson (1984). The full conditional posterior for U is proportional to

$$g(U) f_{\text{Wishart}^{-1}}(U|\Psi_0 + B, r_0 + n - 1),$$

where $f_{\text{Wishart}^{-1}}$ is the inverted Wishart density function. We use this inverted Wishart distribution $W^{-1}(\Psi_0 + B, r_0 + n - 1)$ as a proposal distribution for U. The acceptance probability in this case is given by

$$\min\left\{1, \frac{g(U^*)}{g(U)}\right\},$$

where $W^* = \Phi W^* \Phi + U^*$.

Move types (e) and (f) are essentially the same as move types (c) and (d) with substitution of Θ, V, M, and η for Φ, U, W, and γ, respectively.

Move (g), updating γ and η, can be implemented by forward-filtering, backward-sampling algorithm (West and Harrison 1997) applied to $y_t - \mu$. Model (6) can be rewritten as

$$y_t - \mu = \lambda_t \mathbf{F} + \delta_t \text{ and } \lambda_t = \lambda_{t-1}\mathbf{G} + \rho_t, \tag{9}$$

where $\lambda_t = [\gamma_t \;\; \eta_t]$ is the state vector at time t, $\mathbf{F} = [\begin{smallmatrix} \mathbf{P} \\ \mathbf{I}_{p\times p} \end{smallmatrix}]$, $\mathbf{G}$ is the $(k+p) \times (k+p)$ matrix, $\mathbf{G} = [\begin{smallmatrix} \Phi & \mathbf{0} \\ \mathbf{0} & \Theta \end{smallmatrix}]$, and $\rho_t = [u_t \;\; v_t]$ with variance matrix $\Omega = [\begin{smallmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & V \end{smallmatrix}]$. To sample from the full conditional posterior $\pi(\lambda_1, \lambda_2, \ldots, \lambda_n|\cdots)$, we sequentially simulate the individual vectors $\lambda_n, \lambda_{n-1}, \ldots, \lambda_1$ as follows:

1. Sample λ_n from $N_q(m_n, C_n)$ where m_n and C_n are obtained from the Kalman filtering recurrences

$$m_{t+1} = m_t \mathbf{G} + e_{t+1} K_{t+1},$$

$$e_{t+1} = y_{t+1} - \mu - m_t \mathbf{G}\mathbf{F},$$

$$K_{t+1} = (\Sigma + \mathbf{F}' R_{t+1} \mathbf{F})^{-1} \mathbf{F}' R_{t+1},$$

$$C_{t+1} = R_{t+1} - R_{t+1} \mathbf{F} K_{t+1},$$

$$R_{t+1} = \mathbf{G} C_t \mathbf{G}' + \Omega.$$

2. For each $t = n-1, n-2, \ldots, 1$, sample λ_t from $N_q(h_t, H_t)$ where $h_t = m_t + (\lambda_{t+1} - a_{t+1}) B_t$, $H_t = C_t - B_t' R_{t+1} B_t$, $B_t = R_{t+1}^{-1} \mathbf{G} C_t$, $a_{t+1} = m_t \mathbf{G}$, and λ_{t+1} is the value just sampled.

Note that the likelihood (8) is invariant with respect to changes in scale of Γ or P (even after the identifiability conditions C1 and C2 are taken into account), and the parameters Γ (and so U) and P are identified except for multiplication by a diagonal matrix (consisting of scale constants), that is, we would estimate $\Gamma D^{-1}(D^{-1} U D^{-1})$ and DP unless we use a very precise informative prior. As already mentioned, knowing (estimating) P up to a normalizing constant fulfills the objective of receptor modeling. It also can be shown that a scale constant matrix D (although it is unknown and depends on the initial value of the parameters) does not vary from iteration to iteration within an MCMC run. In this sense, our MCMC scheme is self-consistent, and so the adjustment for the scale constant matrix does not need to be made at each step. If the scale constant (the matrix D) is ever known, the adjustment can be directly applied to the posterior summaries simply by multiplying (or dividing) by D. Care must be taken in specifying the initial values for the parameters or hyperparameters for the prior distributions to ensure that at least they are approximately on the same scale.

Finally, the posterior probability statements can be made directly on the identifiable quantities such as the normalized P or the scaled matrix of U (i.e., the correlation matrix of Γ) as discussed in Besag et al. (1995).

Remark 1. When γ_t and ε_t are assumed to be independent, it can be easily shown that under prior distributions $\gamma_t \sim N_q(0, \Xi_0)$, the full conditional distribution for γ_t, $\pi(\gamma_t|\cdots)$, is a normal distribution through conjugacy, that is,

$$\gamma_t|\cdots \sim N_q((y_t - \mu)\Sigma_\varepsilon^{-1} P'(P\Sigma_\varepsilon^{-1}P' + \Xi_0^{-1})^{-1}, (P\Sigma_\varepsilon^{-1}P' + \Xi_0^{-1})^{-1}),$$

where $\Sigma_\varepsilon = \text{cov}(\varepsilon_t) = \text{diag}(\sigma_{\varepsilon 1}^2, \ldots, \sigma_{\varepsilon p}^2)$. This can be updated using a Gibbs sampler, and with moves (a) and (b) where $Y - \mathbf{1_n} \otimes \mu - \mathrm{H}$ and σ_j^2 are replaced by $Y - \mathbf{1_n} \otimes \mu$ and $\sigma_{\varepsilon j}^2$, respectively, it completes one cycle of MCMC when the observations are treated as independent. In Section 4, this approach is also compared to our time series approach when the observations are actually dependent.

4. SIMULATION

The data are generated by model (3) with $p = 7$, $n = 200$, $q = 3, \sigma_1^2 = \cdots = \sigma_7^2 = 1$, $\phi_1 = \phi_2 = \phi_3 = .8$, $\xi_0 = (10, 12, 14)$, $U = \sigma_u^2 \mathbf{I}_{3\times 3}$ where $\sigma_u^2 = 3$, $\theta_1 = \cdots = \theta_7 = .7$, $V = \sigma_v^2 \cdot \mathbf{I}_{7\times 7}$ where $\sigma_v^2 = 1$. The initial values of α and η are given by $\alpha_{1k} = \xi_0 + \sqrt{\sigma_u^2/(1-\phi_k^2)} Z_k$, where $Z_k \sim N(0, 1)$, $k = 1, 2, 3$, and $\eta_{1j} = \sqrt{\sigma_v^2/(1-\theta_j^2)} Z_j$, $j = 1, \ldots, 7$, respectively. The true source composition matrix P_0 (normalized to sum to 1) is given in Table 2. It follows from (4) and (5) that $W = 8.33 \cdot \mathbf{I}_{3\times 3}$ and $M = 1.96 \cdot \mathbf{I}_{7\times 7}$. This is equivalent to generating the data from model (6) with $\gamma_{1k} = \sqrt{\sigma_v^2/(1-\phi_k^2)} Z_k$ and $\mu_0 = \xi_0 P_0 = (57.96, 54.44, 26.11, 47.15, 29.63, 27.90, 55.63)$. The

Table 2. True Source Composition Profiles (P_0) for Simulated Data

	1	2	3	4	5	6	7
Source 1	0	.248	0	.102	.306	.128	.216
Source 2	.242	0	.266	0	.009	.044	.440
Source 3	.311	.250	.039	.302	0	.099	0

NOTE: Each source profile is normalized to sum to one.

number of sources, q, is assumed to be known throughout the simulation.

In implementing MCMC, we take $\alpha_{0j} = 3$ and $\beta_{0j} = 2$ for the prior on σ_j^2, $j = 1, \ldots, 7$, $r_0 = 10$, and $\Psi_0 = 20 \cdot \mathbf{I}_{3\times 3}$ for the prior on U, and set the degrees of freedom for the prior on V equal to 13 and the scale matrix equal to $6 \cdot \mathbf{I}_{7\times 7}$, each ensuring a proper but relatively diffuse prior. For the nonzero elements of P, the corresponding elements of c_0 and C_0 are set equal to 1 and 1000, respectively, which reflects the lack of information on P. A uniform random matrix with zeros preassigned is used for an initial value of P.

Table 3 contains posterior summaries for some model parameters, based on 5,000 values subsampled from 50,000 iterations following a 50,000 burn-in period. For the source composition matrix P, these summaries are obtained in terms of normalized P (sum to 1) because it is identified only up to a constant multiplier as mentioned in Section 3. The AR coefficients ϕ_k for the source contributions are estimated to be $\hat{\phi}_1 = .817$, $\hat{\phi}_2 = .797$, and $\hat{\phi}_3 = .815$, with the corresponding posterior standard deviations .042, .046, and .042, respectively.

Posterior summaries obtained from the approach assuming independence (see Remark 1) are also given in Table 4. Because this approach does not decompose the error variances into Σ and M, we treat the estimates of the error variances as the estimates for $\Sigma_\varepsilon^2 = \text{diag}(\sigma_{\varepsilon 1}^2, \ldots, \Sigma_{\varepsilon p}^2) = \Sigma + M = 2.96$. The hyperparameters of the priors on $\sigma_{\epsilon j}^2 (j = 1, \ldots, 7)$ and γ_t are taken as $\alpha_{0j} = 2$, $\beta_{0j} = 5$, $j = 1, \ldots, 7$, and $\Xi_0 = 10 \cdot \mathbf{I}_{3\times 3}$, respectively. We use the same prior distribution for P as above. The results are based on a posterior sample of size 5,000 obtained by subsampling from 50,000 values following a 50,000 burn-in period.

Figure 4 shows the side-by-side barplots of the true source compositions (P_0) and the posterior mean of P from two different approaches, time series approach ($\widehat{P_{ts}}$) and approach assuming independence ($\widehat{P_{\text{indep}}}$), with R^2 values between P_0 and estimates. In terms of point estimates, $\widehat{P_{ts}}$ and $\widehat{P_{\text{indep}}}$, there does not seem to be a big difference in this case between the two approaches. However, by comparing Tables 3 and 4, it can be noted that the approach accounting for dependence in the data yields much better results in terms of uncertainty estimates (such as the posterior standard deviations and credible intervals) than the approach not accounting for dependence. In Table 3, only three of the 15 (nonzero) elements of P_0 lie outside the 90% credible intervals (all are within the 99% credible intervals though we do not report them in the table) whereas in Table 4, six elements of P_0 fall outside the 90% credible intervals (five of them are not captured even by the 99% credible intervals). Simultaneous credible regions for the whole matrix P_0 can also be constructed using the method (based on order statistics) suggested in Besag et al. (1995). Table 3 includes

Table 3. Summaries of Posterior Distributions for P, θ, Diagonal Elements of V, and Σ When the Data Is Generated by Model (3) and the Time Series Approach is Used

Parameter	j	1	2	3	4	5	6	7
P_{1j}	Mean	0	.242	0	.102	.308	.129	.220
	SD	0	.008	0	.010	.008	.006	.013
	LSCR	0	.223	0	.077	.290	.115	.187
	LCI	0	.229	0	.086	.296	.120	.199
	UCI	0	.254	0	.117	.321	.138	.241
	USCR	0	.261	0	.124	.328	.143	.252
P_{2j}	Mean	.207*	0	.256	0	.026	.062*	.447
	SD	.018	0	.011	0	.016	.009	.012
	LSCR	.162	0	.231	0	.001	.039	.417
	LCI	.177	0	.241	0	.003	.048	.428
	UCI	.234	0	.276	0	.055	.076	.465
	USCR	.248	0	.285	0	.071	.083	.474
P_{3j}	Mean	.320	.242	.029	.318*	0	.091	0
	SD	.008	.008	.009	.007	0	.006	0
	LSCR	.300	.221	.006	.300	0	.076	0
	LCI	.307	.228	.014	.306	0	.081	0
	UCI	.333	.256	.044	.330	0	.101	0
	USCR	.340	.263	.051	.337	0	.106	0
θ_j	Mean	.654	.779	.697	.600	.512	.582	.503
	SD	.216	.125	.188	.197	.255	.124	.251
V_{jj}	Mean	.750	.892	.713	.910	.850	.810	.925
	SD	.361	.405	.277	.397	.394	.303	.527
σ_j^2	Mean	1.020	.947	1.489	.840	1.622	1.178	1.040
	SD	.448	.321	.382	.337	.520	.321	.545

NOTES: 1. SD stands for the posterior standard deviation; 2. LCI and UCI stand for the lower limit and upper limit of the 90% credible interval; 3. Asterisk (*) indicates that the true parameter value is not captured by the 90% credible interval; 4. LSCR and USCR stand for the lower limit and upper limit of the 80% simultaneous credible region.

Table 4. Summaries of Posterior Distributions for the Parameters P and Σ_ε When the Data Is Generated by Model (3) but the Approach Assuming Independence (given in Remark 1) Is Used

Parameter	*j*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
P_{1j}	Mean	0	.240	0	.101	.305	.130	.224
	SD	0	.005	0	.006	.005	.004	.009
	LSCR	0	.227	0	.086	.293	.121	.203
	LCI	0	.231	0	.091	.297	.124	.210
	UCI	0	.248	0	.111	.314	.136	.238
	USCR	0	.253	0	.116	.318	.139	.245
P_{2j}	Mean	.196*	0	.262	0	.024	.065*	.453
	SD	.012	0	.008	0	.013	.006	.008
	LSCR	.166	0	.242	0	.001	.050	.433
	LCI	.176	0	.249	0	.004	.055	.439
	UCI	.216	0	.275	0	.045	.075	.466
	USCR	.224	0	.282	0	.058	.080	.473
P_{3j}	Mean	.325*	.247	.020*	.318*	0	.090*	0
	SD	.005	.006	.006	.005	0	.004	0
	LSCR	.312	.234	.005	.306	0	.079	0
	LCI	.316	.238	.010	.310	0	.083	0
	UCI	.334	.256	.030	.326	0	.097	0
	USCR	.338	.261	.035	.330	0	.101	0
$\sigma^2_{\varepsilon j} = 2.961$	Mean	2.981	3.312	2.751	2.268	2.449	2.519	3.671
	SD	.782	.557	.497	.523	.744	.272	1.403

NOTES: 1. SD stands for the posterior standard deviation; 2. LCI and UCI stand for the lower limit and upper limit of the 90% credible interval; 3. Asterisk (∗) indicates that the true parameter value is not captured by the 90% credible interval; 4. LSCR and USCR stand for the lower limit and upper limit of the 80% simultaneous credible region.

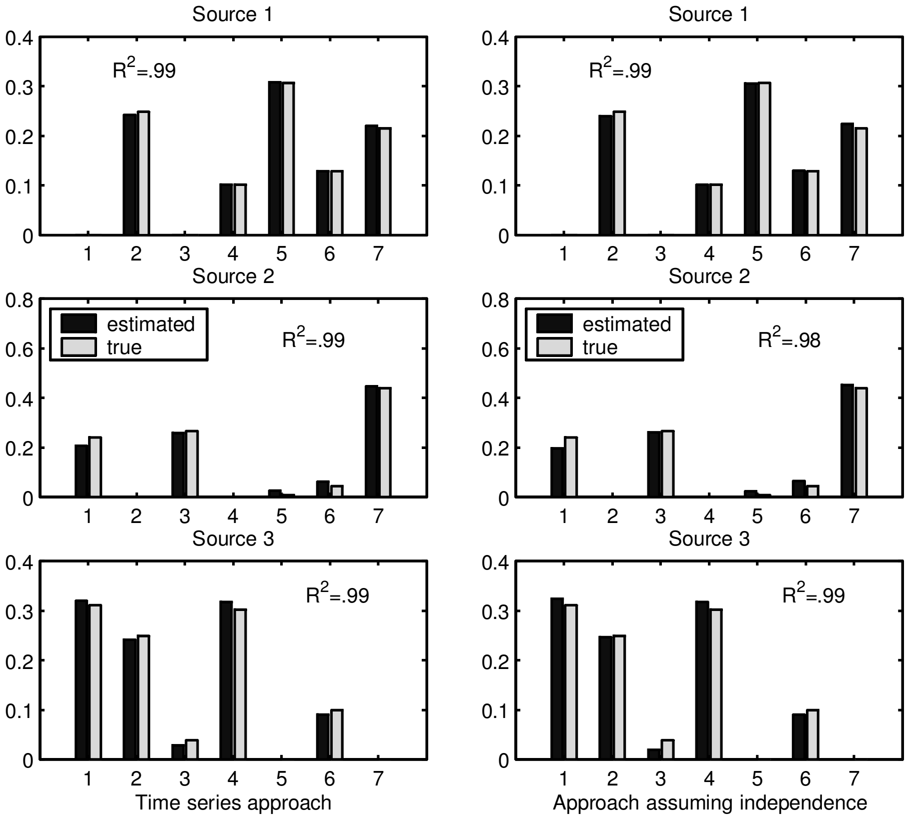

Figure 4. Side-by-side Barplots of the True Source Compositions (P_0) and the Estimated Compositions Obtained From Two Different Approaches, Time Series Approach and Approach Assuming Independence.

the 80% credible regions and these contain all elements of P_0 (the same holds for the 70% credible regions too). In Table 4, five elements of P_0 are still outside the 80% credible regions (they are not captured even by the 90% credible regions). This is a natural consequence of not taking into account the correlation in the errors into the calculation of standard errors (posterior standard deviations here). In fact, the posterior standard deviations in Table 4 are much smaller than they should have been.

To see how much performance of the methods depends on underlying distributional assumptions, a few sensitivity analyses were carried out. To simulate asymmetrically distributed data (lognormallike distribution of species), we generate factors (α) and errors (η and δ) of model (3) from the multivariate lognormal distributions with the parameters $(1.5 \cdot \mathbf{1}_3, .1 \cdot \mathbf{I}_{3\times3})$, $(\mathbf{0}_7, .2 \cdot \mathbf{I}_{7\times7})$, and $(\mathbf{0}_7, .2 \cdot \mathbf{I}_{7\times7})$, respectively. The errors are centered so that they have mean zero before added to $\alpha_t P$ in model (3). The same values for P_0, ϕ_k ($k = 1, \ldots, 3$), and θ_j ($j = 1, \ldots, 7$) as before are used. The proportions of the error standard deviations to the model standard deviations are about 10–30% in this case. The histograms of the resulting simulated data are of lognormal shape. The two approaches, time series approach and the approach assuming independence, developed under the assumption of normally distributed factors and errors, are applied to this data. This enables evaluation of the performance of the two approaches when the distributional assumptions are violated. The performance of the two approaches is not affected by violation of distributional assumptions. Although R^2 values (between P_0 and estimated P) for two approaches are both close to .99, the time series approach outperforms the approach assuming independence in terms of uncertainty estimation. For the time series approach, only one element of P_0 falls outside the corresponding 90% credible interval, and all are captured by the 99% credible intervals. On the other hand, for the approach assuming independence, 8 out of 15 (nonzero) elements of P_0 are outside the 90% credible intervals and 4 are not captured even by the 99% credible intervals.

Next we investigate the effect of serial correlations on uncertainty estimation of P. Several datasets are generated using different values of θ and ϕ (ranging from 0 to .9) under model (3), with an appropriate adjustment to the size of error variances so that the proportions of the error standard deviations to the model standard deviations are approximately in the range of 10–30%. The same P_0 (given in Table 2) is used throughout the simulation. The two methods are then applied to the same dataset each time and their performance is compared in terms of the uncertainty estimates for P. When there is only weak serial correlation in the errors ($\theta_j < .5, j = 1, \ldots, 7$), both methods seem to perform well regardless of the value of ϕ (almost all elements of P_0 lie in the corresponding 90% credible intervals for both approaches). The time series approach outperforms the approach assuming independence when there is moderate serial correlation in the error (θ is greater than .5). As the value of θ increases, performance of the approach assuming independence deteriorates considerably whereas the time series approach works consistently well. When $\theta_j = \phi_k = .9$ ($j = 1, \ldots, 7$, $k = 1, \ldots, 3$), 10 elements of P_0 fall outside the corresponding 90% credible intervals (8 of them are outside the 99% credible intervals) obtained from the approach assuming independence, whereas all elements of P_0 are within the 90% credible intervals obtained from the time series approach. We also observed the same tendency for asymmetrically distributed data generated using lognormally distributed factors and errors with various values of θ and ϕ. Our limited simulation study suggests that inflation of estimated errors using the approach assuming independence would be nonignorable whenever we observe moderate to strong serial correlation (say greater than .5) in the residual plot.

5. APPLICATION TO ATLANTA DATA

The 1990 Atlanta data described in Section 1 has two types of temporal dependence structure, correlation in α and correlation in ε (see Figures 2 and 3). We use model (6) with $q = 3$ to analyze this dataset consisting of 538 measurements on 9 chemical species. For identifiability conditions, zeros are preassigned for CyHx + 2MHx (cyclohexane + 2-methylhexane) and 2,3-dimethylpentane (2,3-DMP) of source 1 (Roadway), acetylene and propene of source 2 (Gasoline), acetylene and 2,3-dimethylpentane (2,3-DMP) of source 3 (Headspace). The information on zero (or near zero) elements was obtained from an environmental engineer's judgment and also from direct source measurements shown in Table 1 (the relative concentrations of those species in each source are observed to be very low).

Our MCMC analysis uses the following hyperparameters for the prior distributions. For the nonzero elements of P, the corresponding elements of c_0 and C_0 are set equal to 1 and 100, respectively, which is a vague (but still proper) specification reflecting the lack of information on the source compositions. For the prior on σ_j^2 ($j = 1, \ldots, 9$), we take $\alpha_{0j} = 5$ and β_{0j} as the jth element of $(8, 2, 20, 8, 4, 4, 2, 2, 8)$, and for the prior on U, we take $r_0 = 20$, $\Psi_0 = \text{diag}(32, 32, 16)$. Finally, for the prior on V, we set the degrees of freedom equal to 20 and the scale matrix equal to $\text{diag}(20, 5, 50, 20, 10, 10, 5, 5, 20)$. This choice of the hyperparameter values was made to ensure that the prior distributions are moderately informative but flexible enough to cover the range of possible values of the parameters. For a starting value of P, a uniform random matrix with zeros preassigned was used (though the chain can converge much faster by the use of good starting values). For each parameter, a posterior sample of size 5,000 was obtained by subsampling every 10th from 50,000 values following a 50,000 burn-in period. We monitored trace plots of all the key parameters, P (normalized), Σ, U (scaled), V, Φ, and Θ, to ensure the chain has converged to the area of high posterior density by the end of the burn-in period. We also inspected the autocorrelation function plots of posterior samples for those parameters though we do not present any of those plots in the article due to limited space. Subsampling every 10th sample seemed to be satisfactory in terms of breaking autocorrelations. When the chain had been further thinned by subsampling every 50th, posterior inference did not seem to change recognizably. Table 5 contains posterior summaries for some model parameters. The AR coefficients ϕ_k for the source contributions are estimated to be $\hat{\phi}_{\text{roadway}} = .718$,

Table 5. Summaries of Posterior Distributions for P, θ, Diagonal Elements of V, and Σ for the Atlanta Data When the Time Series Approach Is Used

Parameter	Species	acetylene	propene	nButane	2Mpentan	3Mpentan	benzene	CyHx+2MHx	2,3-DMP	2,2,4-TMP
	j	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
roadway	Mean	.313	.137	.294	.067	.036	.135	0	0	.019
	SD	.022	.010	.033	.006	.003	.008	0	0	.006
	LSCR	.265	.116	.193	.048	.027	.118	0	0	.003
	LCI	.280	.123	.237	.056	.031	.124	0	0	.008
	UCI	.351	.154	.344	.076	.041	.148	0	0	.028
	USCR	.381	.167	.368	.080	.044	.158	0	0	.035
gasoline	Mean	0	0	.218	.198	.115	.070	.097	.084	.218
	SD	0	0	.036	.007	.004	.004	.007	.006	.014
	LSCR	0	0	.112	.182	.105	.059	.082	.071	.187
	LCI	0	0	.156	.188	.108	.064	.087	.075	.197
	UCI	0	0	.274	.209	.122	.077	.108	.093	.241
	USCR	0	0	.300	.218	.128	.081	.117	.101	.260
headspace	Mean	0	.004	.733	.130	.068	.040	.007	0	.019
	SD	0	.004	.022	.010	.006	.007	.005	0	.010
	LSCR	0	.000	.673	.104	.052	.020	.000	0	.000
	LCI	0	.000	.700	.114	.058	.028	.001	0	.004
	UCI	0	.011	.768	.147	.077	.051	.016	0	.035
	USCR	0	.017	.788	.158	.084	.060	.023	0	.046
θ_j	Mean	.569	.755	.598	.230	.249	.408	.299	.550	.722
	SD	.089	.044	.096	.070	.061	.093	.064	.058	.043
V_{jj}	Mean	.904	.157	3.696	.549	.209	.187	.207	.089	.563
	SD	.203	.033	1.201	.130	.047	.033	.046	.020	.134
σ_j^2	Mean	.711	.088	2.232	.147	.056	.109	.068	.032	.220
	SD	.116	.012	.526	.015	.005	.013	.009	.003	.026

NOTES: 1. SD stands for the posterior standard deviation; 2. LCI and UCI stand for lower limit and upper limit of the 90% credible interval; 3. LSCR and USCR stand for lower limit and upper limit of the 80% simultaneous credible region.

$\hat{\phi}_{\text{gasoline}} = .675$, and $\hat{\phi}_{\text{headspace}} = .282$, with the corresponding posterior standard deviations .024, .027, and .095, respectively.

We also report posterior summaries obtained from the approach assuming independence (Remark 1) in Table 6. For hyperparameters of the prior on $\sigma^2_{\varepsilon j}$ $(j = 1, \ldots, 9)$, we take $\alpha_{0j} = 2$ and β_{0j} as the jth element of $(5, 3, 10, 4, 3, 3, 2, 2, 4)$, and for the prior on γ_t, we set $\Xi_0 = \text{diag}(3, 3, 2)$. For P, the same vague prior distribution as before (elements of c_0 and C_0 are 1 and 100, respectively) is used. The results are based on a posterior sample of size 5,000 obtained by subsampling from 50,000 values following a 50,000 burn-in period.

Because the true source composition P_0 is unknown (as opposed to the simulation study), an objective comparison of the performance of two approaches is difficult. As a guideline, we first compare the estimated source compositions from two approaches with direct source measurements, P_{measured}, given in Table 1. Figure 5 shows the side-by-side barplots of the measured source compositions and estimated compositions from the two different approaches, the time series approach $(\widehat{P}_{ts})$ and the approach assuming independence $(\widehat{P}_{\text{indep}})$, with R^2 values between measured and estimated compositions. General patterns are similar for the source compositions derived from ambient data ($\widehat{P}_{ts}$ and $\widehat{P}_{\text{indep}}$) and for direct source measurements (P_{measured}), which is good for source identification purposes. For the roadway profile, however, there seems to be some deviation between estimated and measured profiles. One possible explanation for this is that the measurements for the roadway profile were taken at the tunnel (during morning rush hour) and might not have captured a full range of variability of all motor vehicle operating conditions in the ambient data.

As mentioned in Section 1, the uncertainty estimates for $\widehat{P}_{\text{indep}}$ provided in Table 6 are considered too small compared to what they should have been (uncertainty estimates for $\widehat{P}_{\text{indep}}$ obtained by adjusting for the correlation). On the other hand, the uncertainty estimates for $\widehat{P}_{ts}$ in Table 5 incorporating the correlation in the data in the estimation procedure are expected to be not only correct but also smaller than the correct uncertainty estimates for $\widehat{P}_{\text{indep}}$ (adjusted for the correlation). In other words, $\widehat{P}_{ts}$ is expected to be more efficient than $\widehat{P}_{\text{indep}}$. For the headspace profile (for which the measured and the estimated compositions show the best agreement for both approaches), 5 elements of the measured headspace composition lie outside the corresponding 90% credible intervals in Table 6, and the same holds even for the 99% credible intervals (not shown in the table) and for the 80% simultaneous credible regions. On the other hand, in Table 5, only 2 elements of the measured headspace profile fall outside the corresponding 90% credible intervals and all are captured by the 99% credible intervals and by the 80% credible regions.

6. CONCLUSIONS AND DISCUSSION

In this article, we have developed a time series extension of multivariate receptor modeling to capture, in the estimation process, extra variability owing to temporal dependence in air pollution data. Recent developments in MCMC methodology make estimation of parameters of complex models possible. By modeling the dependence structure, we

Table 6. *Summaries of the Posterior Distribution for the Parameters P and Σ_ε for the Atlanta Data When the Approach Assuming Independence Is Used*

Parameter	*Species*	*acetylene*	*propene*	*nButane*	*2Mpentan*	*3Mpentan*	*benzene*	*CyHx+2MHx*	*2,3-DMP*	*2,2,4-TMP*
	j	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
roadway	Mean	.309	.132	.287	.075	.041	.134	0	0	.023
	SD	.009	.004	.014	.003	.002	.003	0	0	.004
	LSCR	.287	.122	.248	.066	.036	.125	0	0	.013
	LCI	.295	.126	.262	.069	.037	.128	0	0	.017
	UCI	.325	.139	.309	.080	.044	.139	0	0	.030
	USCR	.334	.143	.322	.083	.046	.143	0	0	.034
gasoline	Mean	0	0	.174	.197	.116	.074	.106	.092	.241
	SD	0	0	.017	.004	.002	.003	.003	.003	.007
	LSCR	0	0	.127	.188	.110	.068	.099	.085	.223
	LCI	0	0	.145	.191	.112	.070	.101	.088	.229
	UCI	0	0	.202	.203	.119	.079	.112	.097	.253
	USCR	0	0	.216	.206	.122	.081	.115	.100	.260
headspace	Mean	0	.002	.611	.183	.101	.053	.034	0	.017
	SD	0	.002	.015	.007	.004	.005	.004	0	.008
	LSCR	0	.000	.571	.167	.092	.041	.023	0	.001
	LCI	0	.000	.585	.172	.095	.045	.027	0	.004
	UCI	0	.006	.635	.194	.107	.061	.040	0	.030
	USCR	0	.010	.648	.202	.112	.066	.044	0	.037
$\Sigma_\varepsilon^2 (= \Sigma + M)$	Mean	1.828	.322	12.498	.117	.051	.174	.158	.048	.662
	SD	.161	.029	.879	.012	.004	.018	.011	.005	.055

NOTES: 1. SD stands for the posterior standard deviation; 2. LCI and UCI stand for lower limit and upper limit of the 90% credible interval; 3. LSCR and USCR stand for lower limit and upper limit of the 80% simultaneous credible region.

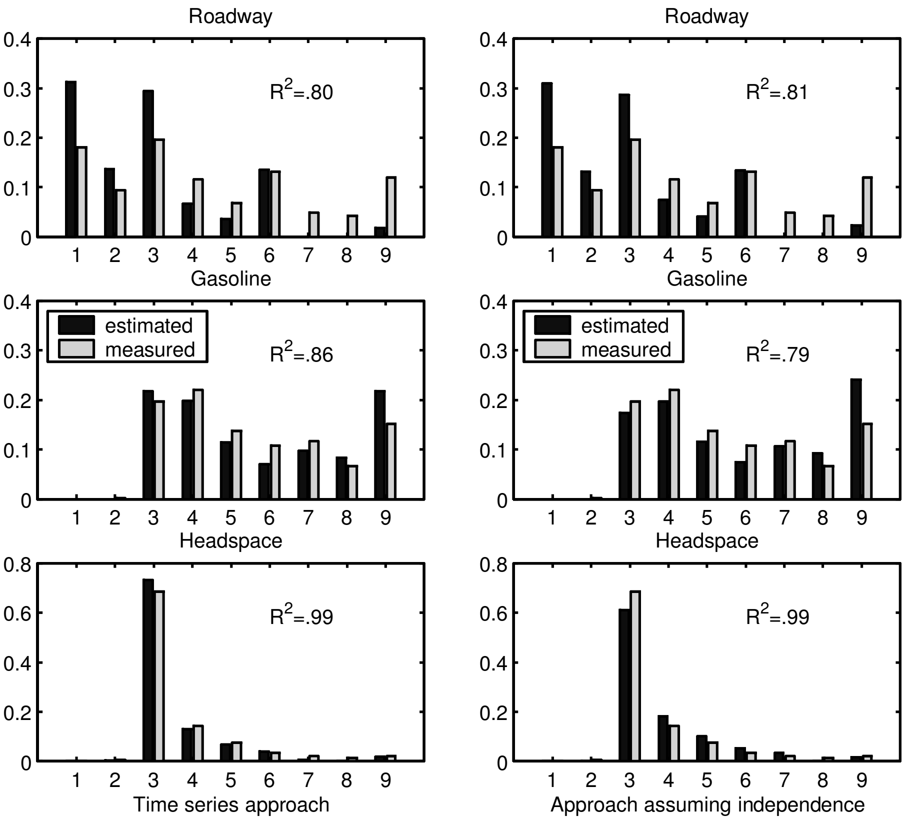

Figure 5. Side-by-side Barplots of the Measured Source Compositions and the Estimated Compositions Obtained From Two Different Approaches, Time Series Approach and Approach Assuming Independence, for the Atlanta Data.

can get more reliable estimates for the source compositions and their uncertainties, which are our primary interest. As a by-product, we can assess the amount of variability and autocorrelation in the source contributions and the errors. It also makes it possible to forecast the level of pollutants (y_{t+k}) and the amount of pollution (α_{t+k}), which has been regarded as one of the model limitations in previous receptor modeling approaches (see the EPA discussion at `http://www.epa.gov/oar/oaqps/pams/analysis/receptor/rectxtsac.html`).

In developing our methods, we assumed that the errors are normally distributed. Environmental data often contain many outliers, and it is sometimes more appropriate to use the lognormal distribution to describe the data even though the results of limited simulations in Section 4 suggest that our methods are robust to violations of normality assumptions. The usual transformation technique does not help in the context of receptor modeling. By log transforming the data the chemical mass balance equation of the model no longer applies directly, and we need to deal with model identifiability using different conditions. Alternatively, we may consider a multivariate T-distribution or a mixture of normal distributions to describe the error distribution. In the application to Atlanta data, the histogram of the residuals for each species looks, in general, bell shaped, but shows a few outliers for some of the species. This might suggest the use of a heavy-tailed distribution for errors even though it was not pursued further in this article. Nonnormal dynamic modeling is still an active research area (see West and Harrison 1997), and we expect that multivariate receptor modeling can be extended further using nonnormal dynamic models.

Air pollution data are often obtained from multiple monitoring sites. When a single pollutant is measured over multiple sites, this can be fitted into the current multivariate receptor modeling framework by treating the different sites as variables. How to incorporate spatial variability as well as temporal variability in modeling when multiple species are measured is a challenging problem. Even in the case of no temporal dependence, this problem remains open.

Finally, chemical reactivity is a major concern that could invalidate a receptor modeling approach because it would violate the basic model assumption of chemical mass balance. For this reason, receptor models have been applied only to relatively unreactive species as mentioned in Section 1. Developing receptor models for reactive species is one of the highest research priorities in air pollution work.

[Received March 2000. Revised June 2001.]

REFERENCES

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: Wiley.

Besag, J., Green, P., Higdon D., and Mengersen K. (1995), "Bayesian Computation and Stochastic Systems," *Statistical Science,* 10, 3–41.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm," *American Statistician,* 49, 331–335.

Conner, T. L., Lonneman, W. A., Seila, R. L. (1995), "Transportation-Related Volatile Hydrocarbon Source Profiles Measured in Atlanta," *Journal of the Air & Waste Management Association*, 45 (5), 383–394.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov chain Monte Carlo in practice*, London: Chapman & Hall.

Gleser, L. J. (1997), "Some Thoughts on Chemical Mass Balance Models," *Chemometrics and Intelligent Laboratory Systems*, 37, 15–22.

Henry, R. C. (1991), "Multivariate Receptor Models," in *Receptor Modeling for Air Quality Management*, ed. P. Hopke, Amsterdam: Elsevier, pp. 117–147.

——— (1997), "History and Fundamentals of Multivariate Air Quality Receptor Models," *Chemometrics and Intelligent Laboratory Systems*, 37, 37–42.

Henry, R. C., and Kim, B. M. (1990), "Extension of Self-Modeling Curve Resolution to Mixtures of More than Three Components. Part 1. Finding the Basic Feasible Region," *Chemometrics and Intelligent Laboratory Systems*, 8, 205–216.

Henry, R. C., Lewis, C. W., and Collins, J. F. (1994), "Vehicle-Related Hydrocarbon Source Compositions from Ambient Data: the Grace/Safer Method," *Environmental Science and Technology,* 28, 823–832.

Henry, R. C., Lewis, C. W., and Hopke, P. K. (1984), "Review of Receptor Model Fundamentals," *Atmospheric Environment,* 18, 1507–1515.

Hopke, P. K. (1985), *Receptor Modeling in Environmental Chemistry,* New York: Wiley.

——— (1991), "An Introduction to Receptor Modeling," *Chemometrics and Intelligent Laboratory Systems*, 10, 21–43.

——— (1997), "The Chemical Mass Balance as a Multivariate Calibration Problem," *Chemometrics and Intelligent Laboratory Systems,* 37, 5–14.

Park, E. S. (1997), "Multivariate Receptor Modeling from a Statistical Science Viewpoint," unpublished Ph.D. dissertation, Texas A&M University, Dept. of Statistics.

Park, E. S., Oh, M. S., and Guttorp, P. (2000), "Multivariate Receptor Models and Model Uncertainty," Technical Report 60, University of Washington, National Research Center for Statistics and the Environment, (http://www.nrcse.washington.edu/research/reports/ papers/trs60_receptor/trs60_receptor.pdf), accepted by *Chemometrics and Intelligent Laboratory Systems*.

Park, E. S., Spiegelman, C. H., and Henry, R. C. (2001), "Bilinear Estimation of Pollution Source Profiles and Amounts by Using Multivariate Receptor Models" (with discussion), *Environmetrics*, to appear.

Press, S. J. (1982), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* (2nd ed). New York: Krieger.

Purdue, L. J. (1991), "The 1990 Atlanta Ozone Precursor Study," in *Proceedings of the Air & Waste Management Association 84th Annual Meeting*, Paper 91-68.8.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics*, 22, 1701–1762.

West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models* (2nd ed.), New York: Springer-Verlag.

Yang, H. (1994), "Confirmatory Factor Analysis and its Application to Receptor Modeling," unpublished Ph.D. dissertation, University of Pittsburgh, Dept. of Mathematics and Statistics.